**Cognizant POV on KP A20 Analytical Digital Platform-**

1.       Use Cases - Data Integration, Data Quality, Data Consumption, Data Governance, Platform

2.       Data Management –

a. Ingestion and Storage - Data Zones (raw/cleansed/curated), Analytical DB’s – Snowflake, Synapse

b.       Compute – Databricks, HDInsight, Hive

c.       Access and discovery - Semantic Layer, APIs enablement, External Tables

3.       Shared Services

a.       Governance – Metadata - Data catalog, Security

b.       Platform – Roadmap Alignment, Interoperability, Maintainability

1. Inferenced from the Evaluation between Snowflake and Synapse (High Level Industry Knowledge, project implementation etc.)
2. Plug in additional technology platforms to meet user needs
   1. Layer 0 / Layer 1/ -- Azure Corporate guidance
   2. Layer 2 – Tenant (Plug and Play) API, Data as a Service, Semantic Layer

**Evaluation Criteria Outline –**

|  |  |
| --- | --- |
| **Primary Criteria** | **Secondary Criteria** |
| **1. Cloud** |  |
|  | **1.1 Cloud Integration** |
|  | **1.2 Elasticity** |
| **2. Performance** |  |
|  | **2.1 Response time & Latency** |
| **3. Storage & Data Distribution** |  |
|  | **3.1 Storage type/formats** |
|  | **3.2 Data Storage** |
| **4. Security, Maintenance & High Availability** |  |
|  | **4.1 Security** |
|  | **4.2 DR / HA** |
|  | **4.3 Maintenance & Work load management** |
| **5. Interoperability/Compatibility & Integration** |  |
|  | **5.1 Native compatibility** |
|  | **5.2 Query capability** |
|  | **5.3 Supported Programing Languages** |
| **6. Cost & Resource availability** |  |
|  | **6.1 Cost** |
|  | **6.2 Resource availability** |
| **7. Vendor** |  |
|  | **7.1 Vendor presence, credibility, training & support** |

A20 Platform Data Management Use Cases

|  |  |
| --- | --- |
| Dataset | Structured and Semi Structured Dataset |
|  | Real Time, Near Time and Batch Ingestion |
| Data Zones | Multi-Tier Data Storage to support Multiple business needs (Raw/Refined/Enriched) |
|  | Data Quality and Enterprise level data conformation |
|  | User Driven Business Data need (Tenant Zone) |
| Data Platform | Scalable and Available Architecture |
|  | Efficient and Performance Oriented Process |

Data Zone Storage (raw, refined, enriched) -

Options:

|  |  |  |  |
| --- | --- | --- | --- |
| Parameters / | Object Store (ADLS) | DB Store (Snowflake) | File Store with Structure (ADLS + Delta Lake) |
| Data Formats Supported | All formats | JSON, CSV, tables, Parquet, ORC … | All formats, Parquet  (Delta Lake are stored as Parquet) |
| Structured | Yes | Yes | Yes |
| Semi Structured | Yes | Yes (Store data as is) , parse using SQL.  Supported format - **JSON, Avro, ORC, Parquet, or XML** | Yes - Schema Enforced – requires parsing and load as Structured in Delta Lake |
| Unstructured | Yes | No | No |
| ACID Transactions | No | Yes (Guarantee) | Yes (Guarantee) |
| Access Methods |  |  |  |
| Connectors | All Azure Services |  |  |
| Security |  |  |  |
| Copy Data | All Formats can be copied | Need Schema | Need Schema |
| Access Control | RBAC, ACL | RBAC, DAC (Discretionary Access Control | RBAC, Table ACL |
| Data Management |  |  |  |
| Data Access |  |  |  |
| Near Real time (>100ms) |  | Yes | Yes |
| Support for Real Time < 100 MS |  |  | Yes |
| Cost | Very low | ?? | ?? |
| Change Data Capture | No Support (Blob Storage can trigger Azure Functions) | Custom Logic | Out of Box |
| Metadata |  |  |  |

Critical Recommendation:

Data Processing (Compute) -

Options:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Data Bricks  ADF + Spark | DB Store (Snowflake) | Synapse Studio (Future) | Informatica PaaS |
| Separation of Storage and Compute | Yes | Yes | TSQL + Spark + Data Management |  |
| Elastic Clustering | Yes | Yes |  |  |
| Workload Isolation | Yes | Yes | Yes |  |
| Multi Cluster Support | Yes | Yes | ?? |  |
| Authentication | Active Directory | single sign-on | Active Directory |  |
| Online Scaling | Yes | Yes | No |  |
| Complex ETL logic support | Yes |  | Yes (Apache Spark) |  |
| GUI/Programming | Databricks – Notebook |  | Synapse Studio – SPARK |  |
| Performance |  |  |  |  |
| Cost |  |  |  |  |
| Priority Workload Schedule |  | No | Yes |  |
| Cashing/In-Memory Processing Support | Yes | Yes | Yes (Preview) |  |
| CI/CD | Native – Azure DevOps | Custom | Native – Azure DevOps |  |
| Metadata |  |  |  |  |
| Data Management | External Integration |  | Seamless Integration |  |
| Data Quality |  |  |  |  |
| Data Lineage |  |  |  |  |
| Metadata Management |  |  |  |  |

Critical Recommendations:

Business Data – Tenant Zone Storage

Options:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Snowflake | Synapse Analytics (SQL DWH New version) | Hive/HDInsight (KP leverage as Analytical query interface) |
| Relational SQL Support | Complete | Complete | Partial |
| ACID Transactions | Yes (Guarantee) | Yes (Guarantee) | Partial |
| DML Support | Yes | Yes | No |
| Compression | Yes | Yes |  |
| Concurrency | Infinite | 128 |  |
| Query Flashback |  |  |  |
| External Tables | Yes (Snowflake External Stage) – Preview Feature | Yes (ADLS/BLOB) |  |
| Materialized Views | Supported (Enterprise Edition) | Supported |  |
| Performance Management and Monitoring tools | Yes | Yes |  |
| Data Migration Complexity | High | High |  |
| Reporting Tool Support | All major Reporting tools – Tableau, PowerBI, etc… | All major Reporting tools – Tableau, PowerBI, etc… |  |
| Statistics collection | Yes | Yes |  |
| Partitioning | Static – Defined during Table creation | Automatically divided into **micro-partitions** |  |
| Data Encryption | Native | Native | User Defined |
| Row/Column Level Security | Enabled with Secure Views | Native Row level , Column (Future) | Enabled with Hive Tables |
| Automatic Statistics Update | Yes | Yes | Manual |
| Metadata |  |  |  |

Critical Recommendations